SPEAKER RECOGNITION USING GMM

G.SUVARNA KUMAR(1) K.A.PRASAD RAJU(2) Dr.Mohan Rao CPVNJ (3) P.Satheesh (4)

1) Sr. Assistant Professor , CSE Department, MVGR College of Engineering , Vizainagaram. emailgsk@gmail.com
2) Student, M.Tech (SE), Avanthi Institute of Engineering & Technology, Makavarapalem, Visakhapatnam. Kaprasad_raju@yahoo.com
3) Professor, Avanthi Institute of Engineering & Technology, Makavarapalem, Visakhapatnam,
4) Associate Professor, CSE Department ,MVGR College of Engineering, Vizainagaram. patchikolla@yahoo.com

Abstract

The idea of the AUDIO SIGNAL PROCESSING (Speaker *Recognition [4] Project) is to implement a recognizer using Matlab which can identify a person by processing his/her voice. The Matlab functions and scripts were all well documented and parameterized in order to be able to use them in the future. The basic goal of our project is to recognize and classify the speeches of different persons. This classification is mainly based on extracting several key features like Mel Frequency Cepstral Coefficients (MFCC [2]) from the speech signals of those persons by using the process of feature extraction using MATLAB. The above features may consists of pitch, amplitude, frequency etc. It can be achieved by using tools like MATLAB. Using a statistical model like Gaussian mixture model (GMM [6]) and features extracted from those speech signals we build a unique identity for each person who enrolled for speaker recognition [4]. Estimation and Maximization algorithm is used, An elegant and powerful method for finding the maximum likelihood solution for a model with latent variables, to test the later speeches against the database of all speakers who enrolled in the database.*

Keywords*:* Speaker Recognition [4], feature extraction, statistical model, Gaussian mixture model, Mel Frequency Cepstral Coefficients, Estimation and Maximization, likelihood.

1. INTRODUCTION

This project encompasses the implementation of a speaker recognition [4] program in Matlab. Speaker recognition [4] systems can be characterised as text-dependent or text-independent. The system we have developed is the latter, text-independent, meaning the system can identify the speaker regardless of what is being said.

The program will contain two functionalities: A training mode, a recognition mode. The training mode will allow the user to record voice and make a feature model of that voice. The recognition mode will use the information that the user has provided in the training mode and attempt to isolate and identify the speaker.

Most of us are aware of the fact that voices of different individuals do not sound alike. This important property of speech-of being speaker dependent-is what enables us to recognize a friend over a telephone. Speech is usable for identification [1] because it is a product of the speaker's individual anatomy and linguistic background. In more specific, the speech signal produced by a given individual is affected by both the organic characteristics of the speaker (in terms of vocal tract geometry [3]) and learned differences due to ethnic or social factors. To consider the above concept as a basic, we have tried to establish an "Speaker Recognition [4] System" by using the simulation software Matlab

Speaker recognition [4] can be classified into identification and verification. *Speaker identification* is the process of determining which registered speaker provides a given utterance. *Speaker verification*, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. The system that we will describe is classified as *text-independent speaker identification* system since its task is to identify the person who speaks regardless of what is saying.

In this paper, we will discuss only the text independent but speaker dependent Speaker Recognition [4] system.

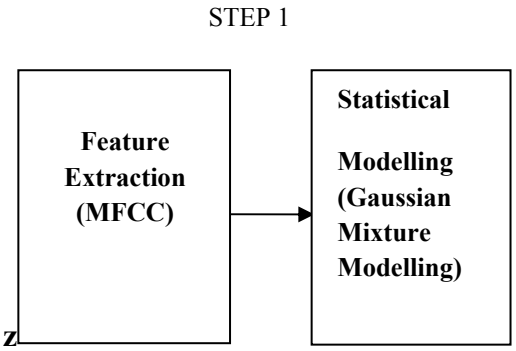
All technologies of speaker recognition [4], identification and verification, text-independent and text-dependent, each has its own advantages and disadvantages and may requires different treatments and techniques. The choice of which technology to use is application-specific. At the highest level, all speaker recognition [4] systems contain two main modules: feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers.

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition [4] task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC [2]). LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue.

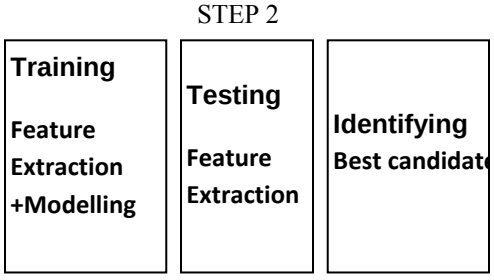
Another popular speech feature representation is known as RASTA-PLP, an acronym for Relative Spectral Transform - Perceptual Linear Prediction. PLP was originally proposed by Hynek Hermansky as a way of warping spectra to minimize the differences between speakers while preserving the important speech information [Herm90]. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency subband in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel e.g. from a telephone line.

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz . MFCC [2] is perhaps the best known and most popular.

Here is just overview of our approach to this project, first we extracted features from the speech signal and then we give them to the statistical model, here we use GMM [6] as statistical model to create a unique voice print for each identity.



Figure(a):Block diagram



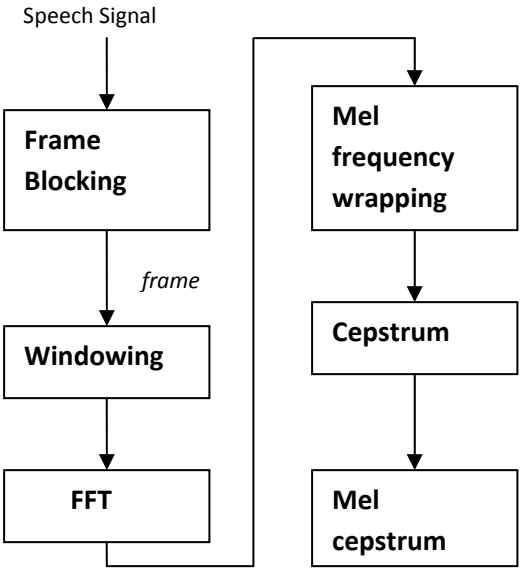
Figure(b):Block diagram

After creation of all voice prints for all identities we check the data base of these voice prints against another voice print which was created by GMM [6] using testing data.

In this project, the GMM [6] approach will be used, due to ease of implementation and high accuracy.

1.1 Mel Frequency Cepstral Coefficients (MFCC's):

MFCC's are coefficients that represent audio, based on perception. It is derived from the Fourier Transform or the Discrete Cosine Transform of the audio clip. The basic difference between the FFT/DCT and the MFCC [2] is that in the MFCC [2], the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT. This allows for better processing of data, for example, in audio compression. The main purpose of the MFCC [2] processor is to mimic the behaviour of the human ears.



Figure(c): MFCC Block Diagram

The MFCC [2] process is subdivided into five phases or blocks. In the frame blocking section, the speech waveform is more or less divided into frames of approximately 30 milliseconds. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The FFT block converts each frame from the time domain to the frequency domain. In the Mel frequency wrapping block, the signal is plotted against the Mel-spectrum to mimic human hearing. Studies have shown that human hearing does not follow the linear scale but rather the Mel-spectrum scale which is a linear spacing below 1000 Hz and logarithmic scaling above 1000 Hz. In the final step, the Mel-spectrum plot is converted back to the time domain by using the following equation:

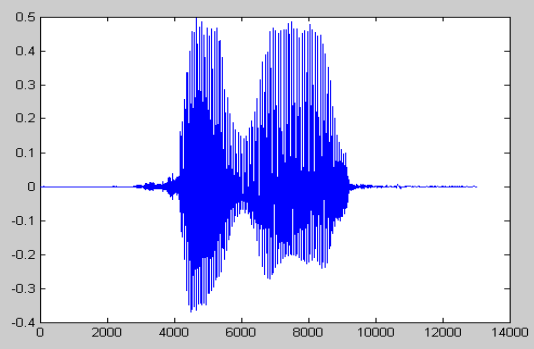
$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700)$$

The resultant matrices are referred to as the Mel-Frequency Cepstrum Coefficients [2]. This spectrum provides a fairly simple but unique representation of the spectral properties of the voice signal which is the key for representing and recognizing the voice characteristics of the speaker.

A speaker voice patterns may exhibit a substantial degree of variance: identical sentences, uttered by the same speaker but at different times, result in a similar, yet different sequence of MFCC [2] matrices. The purpose of speaker modelling is to build a model that can cope with speaker variation in feature space and to create a fairly unique representation of the speaker's characteristics.

1.2 Feature Extraction Module:

Input: Digital speech signal (vector of sampled values)



Figure(d): Sample speech signal.

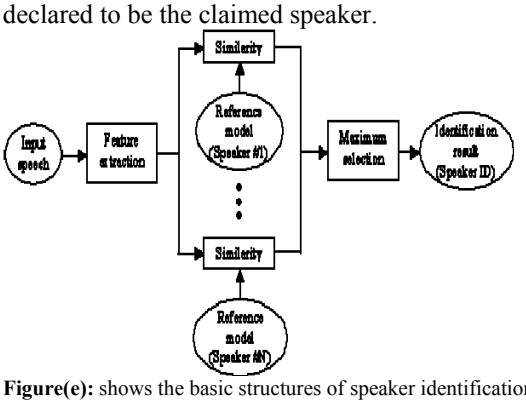
Output: A set of acoustic vectors

In order to produce a set of acoustic vectors, the original vector of sampled values is framed into overlapping blocks. Each block will contain N samples with adjacent frames being separated by M samples where M < N. The first overlap occurs at N-M samples. Since speech signals are quasi stationary between 5msec and 100msec, N will be chosen so that each block is within this length in time. In order to calculate N, the sampling rate needs to be determined. N will also be chosen to be a power of 2 in order to make use of the Fast Fourier Transform in a subsequent stage. M will be chosen to yield a minimum of 50% overlap to ensure that all sampled values are accounted for within at least two blocks. Each block will be windowed to minimize spectral distortion and discontinuities. A Hamming window will be used. The Fast Fourier Transform will then be applied to each windowed block as the beginning of the Mel-Cepstral Transform. After this stage, the spectral coefficients of each block are generated. The Mel Frequency Transform will then be applied to each spectrum to convert the scale to a mel scale. The following approximate transform can be used.

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700)$$

Finally, the Discrete Cosine Transform will be applied to each Mel Spectrum to convert the values back to real values in the time domain.

After creating speaker model we need to identify speaker based on some features such as MFCC [2] as mentioned above. The features of each user are matched against unknown user. And the speaker with best score is declared to be the claimed speaker.



Figure(e): shows the basic structures of speaker identification.

2.MATHEMATICAL BACKGROUND:

2.1 Gaussian mixture probability density function:

After extracting features we need to create a speaker model using some statistical model like GMM [6] statistical model.

Finite mixture models and their typical parameter estimation methods can approximate a wide variety of pdf's and are thus attractive solutions for cases where single function forms, such as a single normal distribution, fail. However, from a practical point of view it is often sound to form the mixture using one predefined distribution type, a basic distribution. Generally the basic distribution function can be of any type, but the multivariate normal distribution, the Gaussian distribution, is undoubtedly one of the most well-known and useful distributions in statistics, playing a predominant role in many areas of applications . For example, in multivariate analysis most of the existing inference procedures have been developed under the assumption of normality and in linear model problems the error vector is often assumed to be normally distributed. In addition to appearing in these areas, the multivariate normal distribution also appears in multiple comparisons, in the studies of dependence of random variables, and in many other related fields. Thus, if there exists no prior knowledge of a pdf of phenomenon, only a general model can be used and the Gaussian distribution is a good candidate due to the enormous research effort in the past. For a more detailed discussion on the theory, properties and analytical results of multivariate normal distributions we refer to.

2.2 Multivariate normal distribution

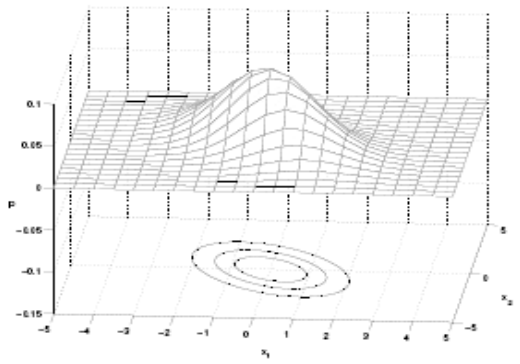
A non-singular multivariate normal distribution of a D dimensional random variable
X →x can be defined as

$$X \sim \mathcal{N}(x;\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[\sim\frac{1}{2}(x\sim\mu)^T\Sigma^{-1}(x\sim\mu)\right]$$

Where μ is the mean vector and ∑ the covariance matrix of the normally distributed random variable X. In Figure 1 an example of 2-dimensional Gaussian pdf is shown.
Multivariate Gaussian pdf's belong to the class of elliptically contoured distributions, which is evident in Fig. 1 where the equi probability surfaces of the Gaussian are centered hyper ellipsoids .
The Gaussian distribution in Eq. 1 can be used to describe a pdf of a real valued random vector (x 2 RD). A similar form can be derived for complex random vectors (x 2 CD) as

$$\mathcal{N}^{\mathbb{C}}(x;\theta,\Sigma) = \frac{1}{\pi^D|\Sigma|}\exp\left[-(x-\theta)^*\Sigma^{*\,1}(x-\theta)\right]$$

Where * denotes adjoint matrix (transpose and complex conjugate).



Figure(f): A two-dimensional Gaussian pdf and contour plots (equiprobability surfaces).

2.3 Finite mixture model

Despite the fact that multivariate Gaussian pdf's have been successfully used to represent features and discriminate between different classes in many practical problems the assumption of single component leads to strict requirements for the phenomenon characteristics: a single basic class which smoothly varies around the class mean. The smooth behavior is not typically the most significant problem but the assumption of unimodality. For multimodality distributed features the unimodality assumption may cause an intolerable error to the estimated pdf and consequently into the discrimination between classes. The error is not caused only by the limited representation power but it may also lead to completely wrong interpretations (e.g. a class represented by two Gaussian distributions and another class between them). In object recognition [4] this can be the case for such a simple thing as a human eye, which is actually an object category instead of a class since visual presence of the eye may include

open eyes, closed eyes, Caucasian eyes, Asian eyes, eyes with glasses, and so on. For a multimodal random variable, whose values are generated by one of several randomly occurring independent sources instead of a single source, a finite mixture model can be used to approximate the true pdf. If the Gaussian form is sufficient for single sources, then a Gaussian mixture model (GMM [6]) can be used in the approximation. It should be noted that this does not necessarily need be the case but GMM [6]s can also approximate many other types of distributions.

The GMM [6] probability density function can be defined as a weighted sum of Gaussians

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c \mathcal{N}(\boldsymbol{x};\boldsymbol{\theta}_c, \Sigma_c)$$

Where c is the weight of c^{th} component. The weight can be interpreted as *a priori* probability that a value of the random variable is generated by the c^{th} source, and thus,

$0 \leq \alpha_c \leq 1$ and $\sum_{c=1}^{C} \alpha_c = 1.$ Now, a Gaussian mixture model probability density function is completely defined by a parameter list [7]

$$\theta = \{\alpha_1, \boldsymbol{\theta}_1, \Sigma_1, \ldots, \alpha_C, \boldsymbol{\theta}_C, \Sigma_C\} \ .$$

2.4. Estimation Maximization:

An elegant and powerful method for finding the maximum likelihood solution for a model with latent variables.
Total data log-likelihood:

$$\mathrm{L} = \ln p(\mathrm{D} \mid \pi, \mu, \mathrm{C})$$

Setting the derivatives of L with respect to the means μ_k to zero, we obtain:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(q_{nk}) \mathbf{x}_n$$

Where

N_k: Effective number of points assigned to the component k

$$N_k = \sum_{n=1}^{N} \gamma(q_{nk})$$

3.ESTIMATION-MAXIMISATION FOR GMM s (Algorithm):

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters.

1. Initialize the means μ_k, covariances C_k and mixing coefficients π_k, and evaluate the initial value of the log likelihood
2. **E step:** Evaluate the responsibilities $\gamma(z_{nk})$ using the current parameter values
3. **M step**: Re-estimate the parameters $\boldsymbol{\mu}_k^{new}$, $\mathbf{C}_k^{new}$, $\boldsymbol{\pi}_k^{new}$ and using the current responsibilities.
4. Evaluate the log likelihood and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

3.1 Log likely hood Calculation:

Another quantity that plays an important role is the conditional probability of **z** given **x**

- Let $\gamma(q_k)$ denote $p(q_k|\mathbf{x})$
- Using Baye's theorem

$$\gamma(q_k) \equiv p(q_k \mid \mathbf{x}) = \frac{p(q_k = 1) p(\mathbf{x} \mid q_{k=1})}{\sum_{j=1}^{K} p(q_j = 1) p(\mathbf{x} \mid q_{j=1})}$$

$$\gamma(q_k) \equiv p(q_k \mid \mathbf{x}) = \frac{\pi_k N(\mathbf{x} \mid \boldsymbol{\mu}_k, \mathrm{C}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x} \mid \boldsymbol{\mu}_j, \mathrm{C}_j)}$$

Where

π_k : prior probability of q_k
$\gamma(q_k)$ is the responsibility that component k takes for 'explaining' the observation x

After the EM step the values converge i.e. they become stable. This is the end of training of speaker models. After this step unknown speaker are tested against the trained samples this is done by using "lmultiguass.m" function.

Start from M initial Gaussian Models N(μk, ∑k),k=1…..M, with equal priors set to P(qk|□)=1/M.

3.2Mathematical Background:

Estimation Step:

Compute the probability P(qk|Xn,□) for each data point Xn to belong to the mixture qk.

$$P(q_k|x_n,\Theta) = \frac{P(q_k|\Theta)\cdot p(x_n|q_k,\Theta)}{p(x_n|\Theta)} = \frac{P(q_k|\Theta)\cdot p(x_n|\mu_k,\Sigma_k)}{\sum_j P(q_j|\Theta)\cdot p(x_n|\mu_j,\Sigma_j)}$$

Maximization Step:

Update means

$$\mu_k^{(new)} = \frac{\sum_{n=1}^{T} x_n P(q_k|x_n,\Theta)}{\sum_{n=1}^{T} P(q_k|x_n,\Theta)}$$

Update Variances

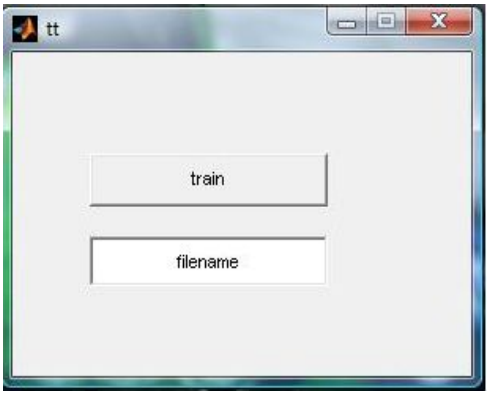
$$\Sigma_k^{(new)} = \frac{\sum_{n=1}^{T} P(q_k|x_n,\Theta)(x_n-\mu_k^{(new)})(x_n-\mu_k^{(new)})^T}{\sum_{n=1}^{T} P(q_k|x_n,\Theta)}$$

Update weights

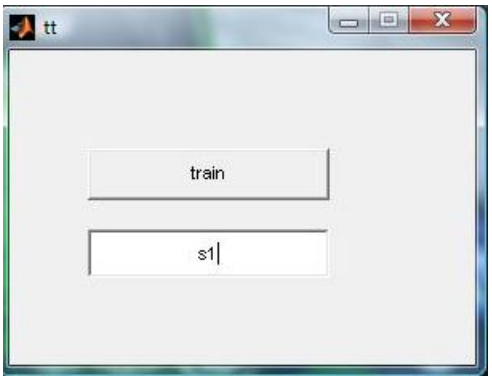
$$P(q_k^{(new)}|\Theta^{(new)}) = \frac{1}{T}\sum_{n=1}^{T} P(q_k|x_n,\Theta)$$

4. RESULTS:

The implementation of this project is done in MATLAB and the results can be seen in a GUI. The GUI takes the filename of the speaker as input and gives the name of the speaker as output. The GUI basically contains a button named "train" when this button is pressed the data is trained and stored in the excel sheets. And it also contains a text field in which the input file name is given. After giving the input file name we have to press "Enter" then the result will be displayed in the same text box. The snapshots of the GUI when providing inputs and when results are displayed are shown below.



Figure(g):GUI snapshot



Figure(h): Input given in GUI

Figure(i): Output in the GUI

5. CONCLUSION:

Over the last decade, the GMM [6] has become established as the standard classifier for text-independent speaker recognition [4]. It operates on atomic levels of speech and can be effective with very small amounts of speaker specific training data. The primary focus of this work was on a task domain for a real application, such as voice mail labelling and retrieval. The Gaussian Mixture speaker model was specifically evaluated for identification tasks using short duration utterances from unconstrained conversational speech, possibly transmitted over noisy telephone channels.

Gaussian mixture models were motivated for modelling speaker identity based on two interpretations. The component Gaussians were first shown to represent characteristic spectral shapes (vocal tract configurations) from the phonetic sounds which comprise a person's voice. By modelling the underlying acoustic classes, the speaker model is better able to model the short term variations of a person's voice, allowing high identification performance

for short utterances. The Gaussian mixture speaker model was also interpreted as a non-parametric, multivariate pdf model, capable of modelling feature distributions.

The experimental evaluation examined several aspects of using Gaussian mixture speaker models for text independent speaker identification. Some observations and conclusions are

- An identification performance of Gaussian mixture speaker model is insensitive to the method of model initialization.
- Variance limiting is important in training to avoid model singularities.
- There appears to be a minimum model order needed to adequately model speakers and achieve good identification performance.
- The Gaussian mixture speaker model maintains high identification performance with increasing population size.

These results indicate that Gaussian mixture models provide a robust speaker representation for the difficult task of speaker recognition [4] using corrupted, unconstrained speech. The models are computationally inexpensive and easily implemented on a real time platform [6].

Furthermore their probabilistic frame-work allows direct integration with speech recognition [4] systems and incorporation of newly developed speech robustness techniques.

6. REFERENCES

[1]J. P. Cambell, Jr., "Speaker Recognition [4]: A Tutorial", *Proceedings of The IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.

[2]Davis and Paul Mermelstein. Steven B. Comparison of parametric
epresentations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[3]Rabiner L, Juang B. H, *Fundamentals of speech recognition* [4]Chap. 2,pp. 11-65, Pearson Education, First Indian Reprint, 2003.

[5] M. Stuttle and M.J.F. Gales, "A Mixture of Gaussians Front End for Speech Recognition [4]," in *Proceedings Eurospeech*, 2001.

[6] J. Gonzalez-Rodriguez, J. Ortega-Garcia, and J.-J. Lucena- Molina, "On the application of the Bayesian approach to real forensic conditions with GMM -based systems," in *2001: A Speaker Odyssey—The Speaker Recognition [4] Workshop*, pp. 135–138, Crete, Greece, June 2001.

[6] D. Reynolds, R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech Audio Process.*, vol. 3, no.1, pp. 72-83, Jan. 1995.